DOCUMENT RESUME

ED 362 027 FL 021 488

AUTHOR Luoma, Sari
TITLE Validating the Certificates of Foreign Language Proficiency: The Usefulness of Qualitative Validation Techniques.
PUB DATE [93]
NOTE 19p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English (Second Language); Foreign Countries; *Language Proficiency; *Language Tests; Second Language Learning; Test Construction; *Test Validity
IDENTIFIERS Finland

ABSTRACT

The Certificates of Foreign Language Proficiency are general purpose tests of language use designed for the adult learner in Finland. This paper deals with the techniques of validation used when designing the test specifications and implementing these into the first versions of the tests. The data comes from the development of tests in one language, English, on two levels: basic and intermediate. The practical purpose of the research presented was to validate the test specifications and tests before their official use. While doing this, several kinds of qualitative validation procedures were used. The purpose of this paper is to examine the relative usefulness of these techniques. Think-aloud protocols and post-test interviews proved the most useful validation tools at this stage of test development. Statistical operations were most useful for examining levels of difficulty, and together with the questionnaires for getting an overall impression of how the testees felt about the test. Think-aloud protocols and post-test interviews were extremely useful and could be recommended as regular tools for test quality maintenance. (JL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Validating the Certificates of Foreign Language Proficiency
The usefulness of qualitative validation techniques

Sari Luoma, Language Centre for Finnish Universities, University of Jyväskylä, Finland

ED 362 027

The Certificates of Foreign Language Proficiency are general purpose tests of language use designed for the adult learner. The paper deals with the techniques of validation used when designing the test specifications and implementing these into the first versions of the tests. The data comes from the development of tests in one language, English, on two levels, Basic and Intermediate.

The practical purpose of the research was to validate the test specifications and tests before their official use. While doing this, several kinds of qualitative validation procedures were used. The purpose of this paper is to examine the relative usefulness of these techniques.

Think-aloud protocols and post-test interviews proved the most useful validation tools at this stage of test development. Statistical operations were most useful for examining levels of difficulty, and together with the questionnaires for getting an overall impression of how the testees felt about the test.

Think-aloud protocols and post-test interviews were extremely useful and could be recommended as regular tools for test quality maintenance.

In this paper I present validation data from a national test development project in Finland. The project is relatively new - it was started in 1992 - so the work I will be presenting is on the initial stages of the validation process, the early piloting period. The aim during the piloting period was to develop valid, functioning test specifications and tests. This involved using several kinds of qualitative validation techniques. We consulted teachers of adults, test specialists and prospective testees along the lines suggested by Kenyon and Stansfield (1993).

Fifteen teachers and administrators received a questionnaire on test specifications, and three testing specialists were consulted on how the operationalisation of the specifications into tests had succeeded. Altogether 124 testees participated in the pilot test versions. From all of them we have 1) test scores and written response data, 2) answers to feedback questions on each task, and 3) answers to questionnaires at the beginning and at the end of the test. The first questionnaire focused on the testees' past experiences in foreign language instruction, testing and language use, and on self-assessment. The post-test questionnaire dealt with how the testees felt about the test, and what they would like to change in it. For investigating the mental processes involved in taking the test, we made 27 post-language-laboratory-test interviews where the testees were allowed to listen to their own performance on tape, and taped three think-aloud protocols on the classroom part of the Intermediate level test.

I will begin this paper by shortly describing the Certificate system and the structure of the tests, and then move on to comparing the usefulness of the different kinds of validation data we were able to gather. I will close the presentation by briefly outlining what kinds of future plans we have for developing the test further.

The test system

The Cer ificates of Foreign Language Proficiency are general purpose tests of language use designed for the adult learner. The target group distinguishes this test from other foreign language tests in Finland, which have traditionally been aimed at children or young people, and have usually been bound with specific educational settings. The development of the test is funded by the Finnish National Board of Education, and the

FL021488

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sari Luoma
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

test is being developed in the University of Jyväskylä by a team of experts in applied linguistics, testing, and educational measurement together with a group of experienced foreign language teachers for adults. The tests we are now developing are aimed at adults who wish to work towards specific goals in their language learning or who need proof of their language skills for their employers in Finland. The certificate system is independent of any educational settings, so it could be attempted by young people as well, but they probably would find some of the topics, situations and roles that appear in the test quite unfamiliar.

The test system of the Certificates of Foreign Language Proficiency consists of tests at three levels: Basic, Intermediate and Advanced. The evaluation of the testees' language proficiency is criterion-referenced and builds on a nine-level scale of proficiency. The scale (Appendix 1) is close to that of the English-Speaking Union's, which was indeed used as one of the models when widening a previous 6-level national scale. The Basic test is intended for skill levels 1-3, the Intermediate test for levels 3-5, and the Advanced test for levels 5-8. We are at present not considering developing a separate test for levels 8-9 because we feel a great degree of tailor-making is required at those levels. This demand, if it should arise, might best be served by a portfolio-type examination combined with some individually designed test tasks.

The languages of the Certificates were initially planned to be English, Swedish, German, French, Spanish and Russian. Recently, Finnish was also included by slightly modifying an already existing test of Finnish as a foreign language. At present, the Basic and Intermediate tests exist in all the languages mentioned, and during the spring of 1993 all have been piloted on a small scale. The development of the Advanced tests was begun in the spring of 1993. The data I present in this paper come from the Basic and Intermediate Certificates in English.

The content specifications of the Certificates consist of definitions of possible *topics, functions, structures, and vocabulary size targets*. These definitions are similar in all the languages of the Certificates, except for the definitions concerning structures, which are language-specific. The topic and function definitions are largely similar to those in international test systems and the Threshold Level, with some changes in emphasis due to the target group, adults and Finns, whose native language and culture differ in several respects from the target languages and their cultures. Thus, work and society are perhaps more strongly stressed than in international language tests while education and learning are somewhat downtoned. Candidates will often find themselves in simulated working conditions, where they receive foreign guests, act as guests or hosts, and communicate on the telephone on business. We are at the moment introducing more systematically a scheme for guiding the operationalisation of the definitions by using the kind of item writers' instructions Davidson and Lynch (1993) present, and the first results look promising.

The theoretical rationale behind the Certificates largely follows Bachman's (1990) view of communicative language ability. As language testers we attempt to concentrate on measuring what Bachman calls 'language competence' while being aware that test takers' personality, strategic competence and knowledge of the world are an integral part of the bits of performance we gather as proof of this competence. We start from Bachman's model because it is comprehensive and was created with specifically testing in mind by a leading expert in language testing. Bachman himself recognises that his model, or

'visual metaphor', cannot capture the interaction which is the essence of communicative language use, but the conceptual diagram is useful for us as test designers. It describes the essential features of what we want to measure, the 'language' part of communicative language ability, and relates the features to each other in a plausible way.

We believe that competencies like those suggested by Bachman underlie all language use. However, what we test is the realization of those competencies in listening, speaking, reading and writing. We also test control of the language system more directly in a subtest on structures and lexis. We do this mainly for diagnostic reasons: this is an effective way of getting feedback on two rather important aspects of learner language. The traditional four skills division is intuitively appealing, and because it seems possible for people to have different skill levels in different skills, we believe the division is motivated. In this respect Milanovic's result that CLA is more clearly divisible at lower proficiency levels and less so at higher levels is interesting, and we are planning to see whether we can find the same phenomenon in our data when we have the appropriate kind of data available.

The Basic and Intermediate tests are, at least at present, very much alike in their structure. Both the Basic and the Intermediate test consist of five subtests: Reading, Writing, Structures and lexis, Listening, and Speaking. On the Basic level, texts are shorter and simpler and, in the case of listening, spoken more clearly, as many (native) speakers would do when talking to a basic level language learner. Production tasks on the Basic level are shorter, more guided and functionally and situationally less demanding than on the Intermediate level, and the structural and lexical items are selected in accordance with the test level.

The subtests in listening and speaking are organised in the language laboratory. The plans for the Advanced level contain an additional test of speaking which would be performed face to face. This was left out of the Basic and Intermediate tests mainly because it would increase the price of a test, but also because of envisaged difficulties in examiner training.

A comparison study between our tests and the equivalent levels of the German ICC test, begun in the summer of 1993, indicates that a better solution would be to have tests of speaking both in the language laboratory and face to face. Some testees feel that a language lab test alone does not give a fully accurate picture of their oral proficiency because talking to a tape recorder is difficult even in the mother tongue: all true interactivity is missing and one cannot use gestures or facial expressions to compensate for flaws or complete the spoken message. Having a natural intonation while talking to a machine is also difficult - anyone who has left a message in an answering machine knows this. The testees also admit, though, that a short, strictly guided face to face test is narrow in a different way, by covering only a very few situations. We are going to pilot a version with both kinds of oral tests in the fall of 1993.

The Reading subtest consists of four short texts, one of which may be modified (eg. shortened), from a range of text types including letters and messages, leaflets, notices, advertisements, articles (newspaper or magazine), reports, instructions and narratives. Starting from these we test skills like understanding gist, understanding specific details, making inferences and deducing meaning from context. Some multiple choice or pairing may be used as task types on one of the texts, but most of the questions are open-ended,

and presented and answered in the mother tongue. The answers are examiner marked. The duration of the subtest in reading is approximately 30 minutes on the Basic level and 45 minutes on the Intermediate level.

The Writing subtest contains two tasks of different levels of formality. The tasks are guided and/or include short bits of resource material. The possible text types required are personal messages, formal and informal letters, reports, instructions and directions. The resource material may in addition to the text types mentioned also be a short article, an ad or a notice. The duration of the subtest is approximately 30 minutes, and it is examiner marked.

The subtest in structures and lexis contains 3-5 tasks which, with the exception of one task, concentrate on the use of structures and lexis in context, both written dialogue and normal writing. We have included written dialogue because we want to focus on everyday structures, lexis and collocations. All the answers, both multiple choice and cloze completions, are at the moment examiner marked. The duration of the subtest is approximately 30 minutes - this may be increased to 45 minutes in the near future.

The Listening subtest has two or three tasks, each of which may consist of several shorter spoken texts from a range of text types including announcements, advertisements, taped messages, (bits of) conversations, and radio broadcasts (factual, narrative, or informative). Some of the listening material is heard twice, some once. The questions are as a rule open-ended, presented and answered in the mother tongue. As in the case of reading comprehension, the use of the target language in questions and answers will possibly be taken up on the Intermediate level. The duration of the Listening subtest is 20-25 minutes.

The Speaking subtest immediately follows the Listening part. After a short warm-up exercise, there are 3-5 tasks chosen from the following: reading out loud, simulated conversations, answering questions about oneself, expressing one's opinions and/or likes and dislikes, reactions in situations, completing utterances and presenting a longer monologue type description, narrative etc. with the help of a cartoon / visual prompt or a presentation skeleton / written prompt. The candidates' oral production is recorded on cassette and afterwards marked by an examiner. The duration of the subtest in speaking is approximately 30 minutes.

The envisaged face-to-face speaking test concentrates on the feature most acutely missing from the test structure at present, genuine interaction. The setting will most likely be paired candidates interacting with a non-native tester. (Native speaker testers will undoubtedly be used when possible, but given that this will be a national system in Finland, it is unrealistic to expect that a sufficient amount of trained native speakers would be available on certain dates everywhere around the country.) The subtest will begin with a warm-up of introductory questions with the tester as the leading character, possibly followed by two role plays where the tester only sets the scene and the bulk of the interaction is between the testees. Then a choice of three topics is presented to the testees, who agree on which topic they want to select. Once selected, the testees will read a short passage or hear an introduction into the topic, and then discuss it with each other. The tester only participates if one of the candidates does not seem to get enough chances to speak, and at the end in order to bring the discussion to a close. The face-to-face test would last approximately 15 minutes.

The evaluation of the candidates' communicative ability is examiner intensive. During the piloting period we found that marking requires 25-35 minutes per candidate once consensus has been reached on which answers are acceptable. The decisions on how the marking will be done in practice have not been made yet, but we are considering training teacher examiners for initial marking and a 10-20 per cent double marking by a core group of examiners.

The certificates awarded for passing the tests report a general mark and a profile for each of the subtests: reading, writing, structures and vocabulary, listening, and speaking. The profile is also given in written form so as to give the candidates enough feedback for guiding their future studies. Giving an evaluation profile is motivated because according to our experience individuals have different needs profiles and different skill profiles. We decided against relying on profile evaluation alone because we believe that we as testers should be able to "see the forest for the trees" and report general proficiency. Also, some employers only seem to be interested in an overall estimate of language proficiency. When asked, most people tend to give a very global assessment of their skill, and many can give a fairly reliable estimate their own general proficiency in all the languages they know (although they will often want to know who wants to hear their estimate and how it is going to be used), and people often find themselves evaluating the proficiency of their (non-native) conversation partners in overall terms. As language testers, we feel that we should be able to do the same but more reliably, and after considering each candidate's speaking, listening, reading, writing, and lexical and grammatical knowledge, we suggest that our estimate is not purely impressionistic but actually fairly systematic.

The findings from the pilot tests

At present we have, all in all, data from 19 candidates on the Basic level Certificate in English and 105 candidates on the Intermediate level. (Recruiting test candidates for the Basic level test in English has been a problem, while candidates for the Intermediate level have been plentiful. In some other languages of the Certificates, most notably Spanish, the situation has been the reverse.) From all the candidates we have three kinds of data: scores and performances on the test tasks, answers to questionnaires given before and after the test, and answers to feedback statements on the test tasks filled in immediately after each task. We also have, so far, three think-aloud protocols on the Intermediate level test, written section, and 17 Intermediate level and 10 Basic level retrospective interviews made after the language laboratory section.

In contrast to multiple choice tests, the interpretation of the statistical analysis of score data on this test also requires response analysis, most of the answer alternatives being determined by the testees. The questionnaire in the beginning of the test gathered information on language education, language use and self-assessment. The questionnaire at the end of the test focused on what the testees thought of the test as a whole and how well they did in it. The feedback statements after each test task were responded to on a 5-point Likert scale (strongly agree - don't know - strongly disagree). The statements were:

I understood what I had to do
There was enough time
The task was easy

I did well on this task
I felt that this task measured my language skills
I liked this task.

In addition, we asked whether the testees were familiar with the task type and if so, from what context, and gave some space for further comments on the task. In the think-aloud protocols, the testees were asked to let as much of their thought process as possible come out on tape. The language was to be the one they thought in. In the retrospective interviews after the language lab, the testees were allowed to listen to the texts they had just worked on again, and also to listen to their own performance on the speaking items. The questions focused on what the testees did while completing the test, how they felt about the tasks and what they would like to change in the tasks.

Simple numerical analysis of the scores, that is, descriptive statistics and correlations, provided some information on which items were good and which needed to be revised. Response analysis gave further light on the nature of some of the problems. The feedback statements gave every testee a chance to say what they thought of the tasks, and they were specifically useful for getting comments on item types, and in the language lab on the appropriacy of the response times. On the other hand, some of those interviewed commented that at least their own comments on the facility of the task and on how they did on it were often too negative, maybe because of a lingering feeling of failure when they did not know one or two words, or were not completely sure about one or two answers. If they had answered the feedback questions after completing the whole test, they might have given themselves more credit. The protocols and the interviews were mainly collected for trying to find out what kinds of processes our tasks made the testees go through, but we were also able to use them for explaining how the items worked.

The adequacy of different validation techniques

Test specifications - questionnaire

The test specifications were created on the basis of recent needs analysis results, existing test specifications, study books and the Threshold Level. Topics and Functions were chosen as the areas to be defined because we wanted to define what the test takers should be able to DO with language in situations which they might encounter and have to manage in the target language. The definitions were made so that they could also be used as a basis for developing new syllabi or modifying existing ones.

After the specifications were drafted and revised in accordance with expert comments, they were sent out in questionnaire format to 15 teachers and administrators for further comments. The commentators gave their opinion on the importance (very important - important - not important) and level (basic - intermediate - too difficult) of each suggested subheading of the topic and function lists. In addition, after each group of topics and functions there was space provided for suggestions on what to change in the list. Nine of the 12 teachers and one of the three administrators returned the questionnaire. One administrator sent back a note saying it was impossible for him to comment on the specifications before seeing an actual test.

This "pre-validation" of the test specifications showed that the specifications were by and large acceptable to teachers of adults, and that we had in some cases been perhaps too specific (some subheadings could be combined). On the basis of this round of comments, Geography and specific characteristics of the Finnish and the target-language countries was taken up as a topic, and culture as a topic was strengthened through adding more specifications on what the topic covered. The questionnaire format was found relatively useful for getting feedback - the return rate might have been lower if we had only had open questions, and the answers would in all likelihood have been less specific, but difficulties with the format were of course also encountered. The most common comments were that the commentators were not quite sure what this or that item on the list referred to, and that it was difficult to tick one alternative when reacting on how important a topic was for tests at two different levels. Only two commentators had found a way around this by making two tick-marks and explaining in the open answers.

The actual validation of the test specifications in relation to draft tests has only started in the summer of 1993. We are trying to go back from test tasks to topics and functions and then to evaluate whether the test adequately covers the specifications, and indeed whether we have defined the correct kinds of descriptors for specifying the test content. For this, the format will have to be a more open-ended questionnaire, followed up by a feedback discussion. Thus at least some of the possible misunderstandings can be clarified and further information on the success of the operationalisation can be gathered.

Pilot test data - descriptive statistics and correlations

Simple statistical operations - descriptive statistics and correlations - worked well for getting an overview of the test and finding out levels of difficulty. The levels of difficulty of each task are important because both our tests attempt to measure where on a range of proficiency levels the testees are. In the beginning, it was difficult to place the tasks exactly; we knew approximately what we were doing but the criterial points were not very clear. This was especially the problem of comprehension tasks, which shows clearly in the descriptive statistics. There were some tasks where the means were very close to the maximum score and standard deviations quite small. (e.g. two tasks where max. was 6 pt, means were 5.36 and 4.76, and std's 1.05 and 1.28, respectively.) We then had to go back to the texts and see what it was that was so easy in them. We concluded that the texts were short and contained no or very few "difficult" words, that they had clear structure and purpose (informative, mostly), and that the questions were of the type "what does the text say about X". (The more difficult texts, by contrast, were longer and contained more "difficult" words, and the tasks required both explanation of text content and interpretation or summarizing.) The tasks that seemed to discriminate better among the test population were naturally also analysed for what makes them more difficult, and whether they indeed were more difficult or just badly constructed.

Table 1 gives some descriptive values for the biggest subsample in the written section of the Intermediate tests, N = 58. The far right column gives the maximum amount of points available for that task. According to the statistics shown, Reading task 4 seems to work quite nicely, with mean at 68% and a 30% standard deviation, while Reading task 5 is far too easy, with mean at 99%. The low standard deviation for Writing task 1 was expected: the evaluation scale ranged from 1 to 5, but because the test was for levels 3-5,

we expected most of the values to be within the smaller range. Closer analysis, prompted by the relatively low mean for the idioms task, revealed that some items in the task were too difficult for the test level.

Table 1. Sample of descriptive statistics

Number of Valid Observations (Listwise) =			58.00		
Task	Mean	Std	Min	Max	Taskmax
Reading 4	5.46	2.42	0.0	8.0	8
Reading 5	3.95	.29	2.0	4.0	4
Writing 1	3.63	.62	2.0	5.0	5
Structures 1	19.19	4.50	5.0	27.0	27
Idioms	4.52	1.57	1.0	7.0	8
Lexis	30.59	4.10	22.0	40.0	40

Correlations were the only kind of statistics possible for analysing such small populations as ours -the group sizes ranged from 5 (Basic level written section) to 88 (Intermediate level lab section). With group sizes of 30 or more, there were many significant and very significant correlations (at $p <= .01$ and .001, respectively). Between subtests, correlations ranged from around .4 (lowest between reading and speaking, and writing and listening) to .6 - .7 (between reading and writing, and structures and lexis with everything). Within subtests, correlations between the different tasks were usually in the .7 to .9 range, with some exceptions. The exceptions included the too easy tasks, which hardly correlated with anything except each other, and one task on collocations and idioms (monolingual, 5-alternative mc). This task was very difficult for the Intermediate level testees: they typically scored 3 out of 8, and the score only correlated significantly with two tasks, one writing task and one structure task, both at the .55 level. The collocations task is obviously too difficult for the testees, but we want to keep the task at least for the time being, in the hopes of getting washback and raising the testees' consciousness of this aspect of foreign language use. Revising the items is necessary, though. The low correlation between tasks of listening and speaking may be, in part, a result of too easy, poor or unclear listening tasks. Another angle on the low correlations is that some of the listening tasks seem to have been uniformly too easy or too difficult, i.e. standard deviations were much lower on these tasks than on the tasks at the more appropriate difficulty level. The listening comprehension subtest as a whole needs further analysis.

The dilemma we are left with is the same that testers have met before: does it just happen to be so that knowledge of grammar correlates well with ability to communicate on the basic and intermediary levels, or are we in spite of conscious attempts to evaluate communication still largely basing our ratings on grammatical accuracy? Or, from the non-native evaluators' point of view, how can we clearly define the line between comprehensibility and incomprehensibility without using grammatical accuracy as the criterion? The relatively high correlations of the structures and lexis tasks with speaking and writing can at least partially be explained by that accuracy and range of expression are among the evaluation criteria for these tasks, and whether the evaluators can

distinguish between 'pure' grammar and grammar from the point of view of communication remains to be seen.

Pilot test data - test takers' perceptions of the tasks

The set of questions after each task (Figure 1 below), concentrating on the perceived quality of each task, provided some information, but their use was not unproblematic. In the final questionnaire 60% of the testees indicated that answering them had affected their concentration somewhat or considerably, and some testees had simply left the feedback questions unanswered. The most useful comments were gathered on timing in the language laboratory. The testees avoid using the end points of the scale, except regarding one point, I understood what I had to do, where they were almost unanimous in agreeing that they did indeed understand. This was probably because the instructions were given in Finnish. Extreme negative answers were most common on the point "I did well on this task". This is not uncommon for Finns, even young Finnish pupils have been proved to rate themselves more strictly than other nations.

Figure 1. Set of questions answered after each task

	strongly agree		don't know		strongly disagree
I understood what I had to do	1	2	3	4	5
There was enough time	1	2	3	4	5
The task was easy	1	2	3	4	5
I did well on this task	1	2	3	4	5
I felt that this task measured my language skills	1	2	3	4	5
I liked this task.	1	2	3	4	5

Have you encountered this task type before? ___ no
___ yes Where? _____

Other comments on the task:

_____

_____

On the think-aloud protocols, it appeared that perceptions of the tasks varied very much, and that the testees chose their answers based on highly different grounds. Answers on whether a testee liked one of the reading tasks were answered: (original in Finnish, translation mine) "I liked this task, well why not, it was so easy", "Why wouldn't I have liked it, it was interesting", "I don't know now I don't really care for reading comprehension, I'll put strongly disagree". The last testee had chosen 'agree' on the previous reading comprehension task regarding the same statement. One could guess that he was perhaps growing tired of the RC part of the test, but perhaps not that he did not do well on the task, because in the next task he clearly reacts negatively because he feels he has failed and says so:"I didn't like this task because I didn't know the word *carpet*". One of the testees commented in the retrospective interview after the lab section that the affective reactions she had indicated directly upon completing a task were perhaps too strong, after a time she might have felt better about her performance. A better place for these questions might thus be the post-test questionnaire. On the other hand, the testees would then have to be reminded of the tasks, and yet they might not remember

what the tasks actually were like. Also, answering a long series of questions at the end of a fairly long test may be demotivating and result in a lot of missing data. One way, provided that there are sufficient numbers of test takers, would be to rotate the questions. One could also collect answers both during and after the test, and then compare the results. This would require some effort from the testees, but, on the other hand, they know that they are taking a pilot test, and probably understand why such feedback questions are being asked.

Each question series on perceptions of the task closed with space for further comments. Some of these focused on sound and voice quality on the language laboratory tape. Those who commented on this usually reacted negatively towards the clarity or heard attitude of some speakers ("mumbly pronunciation", "unclear", "too slow and deliberate", "depressing, unenthusiastic", and "irritating s-sound") or towards that at some points one could hear from the intonation of the speakers that they were reading the dialogue. Some testees were also disturbed by some dialectal vowel qualities of the instructions reader in Finnish. Especially those who did well on the test had time to notice these features. This was a pilot test, but that is not a good enough excuse. The conclusion to be drawn is that in order to help all testees concentrate on the content of the tape, the format has to be very neutral in all possible ways.

Pilot test data - The pre- and post-test questionnaires

The background information gathered in the pre-test questionnaire shows that our pilot testees represent the target population of the test fairly well. Of the 124 participants, 86 were female and 38 male. In age they ranged from about 20 to about 50; a few older people also attended. 40% had college background, 30% were vocationally trained, 10% had university degrees and the remaining 20% had been trained by their employers. 30% had never been to a country where English was used as the main medium of communication, 40% had spent a week, 20% a month and 10% a year in an English-speaking country. Most of the testees had had four to ten years of English during their schooling, but for many of the participants this was more than ten years ago. About 60% had had English as a hobby for a year or two, but for some up to five years. Most (around 70%) had used at least some English at work and during freetime, for instance while telling the way to foreigners, and listening to the soundtrack on tv and the movies. 50% also said they used English with friends in Finland. Over 90% had used English when travel8ling, though only 25% had used English abroad for professional purposes. All of this is more or less expected, given that the levels tested were Basic and Intermediate.

The testees assessed the level of their language skills on a slightly modified version of the scale given in Appendix 1. The descriptions are somewhat ambiguous, which may explain why only about half of them estimated themselves to be on the level where our pilot test placed them. All the correlations were highly significant ($p < .001$), but values were low, between .41 (writing) and .58 (overall). Moreover, overestimations were slightly more common than underestimations, which is contrary to the normal result in studies of Finns. Many testees commented in the final questionnaire that they had thought the test would be easier than it was - this was in part due to some changes in the test after the advertisement for the pilots had come out.

As a part of the self-assessment, the testees were asked which skill was best and which

poorest developed in their English. 45% estimated they were strongest in reading, 30% felt their strongest skill was listening. 45% believed that their weakest side was speaking, and 30% that they did worst at writing. The reason for such a big portion of people believing that speaking is their weakest developed skill is most likely the age of the target population. Most of these testees never practised speaking at school, or were indeed discouraged from speaking unless they could express themselves without grammatical flaws.

The final questionnaire focused on feedback from the test. More than 90% said they liked the test, and 80% of the people were of the opinion that the test corresponded fairly well or very well with real-life communication. 50% thought they got enough information about the test. Those who wanted more mentioned wanting to know about duration, task types and skill levels tested. Sample items were also brought up. We are in the process of preparing a test booklet for each test level in each language, so this problem will be avoided in the future.

The questionnaires were effective in collecting this type of background information and relatively limited affective reactions and self-assessment. The questionnaires were kept as short as possible, but still amounted to about ten minutes in the beginning and five minutes at the end. Some testees obviously felt this was too long, because they simply left some of the questions unanswered. Thus, if we want to consider lengthening the final questionnaire to collect task-by-task feedback, we will most probably have to devise multiple choice questions in order to get any answers at all.

Pilot test data - The protocols

When gathering the think-aloud protocols, we asked the testees to express as much of their thought process as they could, in the language they used when thinking. We asked them to also indicate reading by reading out loud a word here, a word there. During the actual think-aloud process, the testee was alone in a room with the test and a tape recorder. We were lucky in finding three "good informants" for this type of data collection; a fourth testee who also volunteered for think-aloud found herself unable to express her thoughts and used self-report instead; this was completed with a retrospective interview on the following day.

Our testees, all intermediate-level candidates, introspected mostly in Finnish. This is at least in part due to the design of the test - the instructions and the questions are in Finnish. The instructions were written so that they made the testees fairly test-conscious, which also shows in their behavior: e.g, as instructed, they read the questions before reading the text. Upon completing a task, they quickly answered the feedback questions and moved on. However, the fact that the questions were there may have made them, as well as all the other pilot testees, more test-conscious.

In reading comprehension, the three used English only for reading the texts. According to the test constructors, the questions test comprehension of main ideas and specific details, and thus require (careful) reading of the question, finding and understanding the answer in the text, and reporting the answer in Finnish. In answering the questions, the testees, after reading the question, usually located the answer first, and then formulated the answer in Finnish: (Original in two languages, translation mine, sections in italics in

English in the original) "According to the text, what is common for Spain, Italy and Turkey in development of taxation? / Oh yea, there, *we had* / right there, they have raised, they have raised taxes enormously". They sometimes indicate uncertainty by saying something like "I wonder if it was so", and usually deal with it by reading the English text again and then answering as they planned the first time. They notice difficult words, which are often also difficult to pronounce, but usually successfully use the context to infer their approximate meaning. One text, a notification in a paper to the effect that the paper does not take responsibility for personal damage caused by answering an ad printed there, was specifically chosen to see whether the testees can do this, the questions focused on main points of the text. All three testees read the text through slowly at the first time, then quickly glanced through it again and then somehow, probably based on holistic impression, inferred the core meaning. A contrary example, and a surprise to the test constructors, there was an individual important vocabulary item in one text which was unfamiliar to many testees (all the three introspecting testees included), the word *carpet*. From the context they could infer that it is a surface material but they did not know whether it is on the wall or on the floor. Answering, one of the testees optimizes his answer. First he tries to get the exact meaning of the word: he repeats the word several times and tries to hear whether it is carpet or tapestry, then he re-reads the passage, and finally decides to use the word 'surface material' in the answer in order to gain the points. He may have realized it did not matter so much whether he knew the exact word, he understood the main idea, which was what the question required.

The reading comprehension subtest appeared in the light of the protocols to be measuring what we wanted it to measure: reading relatively easy, everyday texts for gist or for detailed comprehension of important points. We were especially happy for the way in which the testees seemed to have to think of and formulate the answer to the question themselves instead of choosing from given alternatives. If we had used multiple choice, the texts would probably have had to be longer and chosen on different grounds, and the item writing process would have been different. The task types used in reading comprehension thus appeared suitable, although it was noted that some texts were too short and easy, and that some more questions could focus on summarizing, drawing conclusions and distinguishing writer attitude. Texts should then be chosen so that such questions can be written.

According to the test constructors, one of the writing tasks required passing information informally (the main content of the message to be written was given in Finnish), the other the expression of own opinions a little more formally, (a letter to the editor on Sunday shopping, based on a short article). In the first task, they expected the testees to review the given content, perhaps imagining themselves in the situation, compose the message, and possibly check it through. In the second, they expected the testees to read the material through, compose the letter and check it for mistakes and content clarity. The protocols show that in both writing tasks, the testees process through Finnish, and often stop to wonder whether some phrase or other that they are using is correct English. In doing that, two of them repeat what they have written over and over, and try to hear whether it is correct, and one reads the phrases through to see whether they look right. All three seem to have trouble with expressing time: (translation mine) "If I put *fifty past eight* then it's *ten to nine / at* well now, in fact I'd write, I'm not sure about those / how they go / I'll put *at twenty fifty* / and I'll add on *o'clock* / but now that's not / I really can't do that / and it looks like French now / so I'll take it away and put it like that...".

One of the testees uses a previous text as a model for his message several times. He also observes his test behavior closely on many levels: how formal/polite he may appear, whether all he writes is functionally necessary in a short note like that, how his handwriting changes during the writing of the message. Only one of the three think-aloud testees checked his production as a whole.

The task types and difficulty seemed to work well for the levels intended. In the writing part, it appears extremely important to consider what we ask the testees to do, and how we ask them to do it. They may or may not pay attention to the frame we have created for the task, but they definitely check that they cover the content we ask them to, and also pay attention to formal correctness. Instructions are important because we can only expect the testees to do as much as, or a little less than, we ask them to. Based on the protocols and response analysis of other testees, it seems that guidance as to required text length and style level might be useful, and a reminder of appropriate textual features like opening and closing a message could also be considered. Without indications of required text length, five short content points were found to be the absolute minimum for the guided task for getting texts long enough to be rated directly on the overall scale. Most of the testees expressed the content asked in the shortest way possible, maybe because of experienced time pressure.

One of the structure tasks was cloze completion with clues in Finnish, the other mc-cloze with alternatives in English. Both tasks had short written dialogue (three to six turns) as texts. The test constructors expected the testees to work exchange by exchange, first reading it through to get the gist, and then completing the dialogue, in the easiest cases without the help of the Finnish clues. The testees did this, but very rarely left the clue in Finnish unused. Indeed, all of them seem to translate surprisingly much word for word and then judge the acceptability of the result: ".. [reads]*I'm pleased with what I'm doing now but things may change in some* / [writes] *some / years y-e-ars / years* / looks funny I think there's something / but *some years* is in a few years [the given clue in Finnish] / looks weird / ..." "[reads] *Now that the two Germanies are united what will they call themselves?* / [writes] *what will / they / they call themselves will they call* what name will they *call* / is that it? / *call / what will they name* / call a name / *call* / I wonder which it is / *call* is to use the telephone / *name* / a name call a name they / *what will they name themselves* / It doesn't sound good / *call* it is that's what it'll be..." Even if the testees are fairly sure their answer is correct, they check it by reading the completed exchange through once more.

The testees are positive towards the use of written dialogue in structure tasks because it brings the tasks closer to real life. The only problem that the test constructors could envisage was that the testees could have felt less demand for accuracy in the conversations where content obviously also counts, but the three candidates showed this was not a serious concern. In the structure tasks the testees obviously had to use what structural knowledge they had, so it looks as if we would be measuring the "right thing". However, the testees' frequent use of Finnish as a basis for choosing the English forms creates a conflict which needs to be considered. At least outside the test situation, we want to encourage monolingual processing in the testees' language use, so why do we use a test form which obviously encourages bilingual processing? We obviously need to experiment with some kind of a monolingual structure task to see whether seeking support from the mother tongue has to do with the accuracy demands of the test situation or whether it is just the bilingualism of the task that causes the language

switching, and then consider which task types are the most appropriate to use in the test.

The lexis tasks were multiple choice. In the collocations tasks the alternatives were given in English, while in the other, out-of-context vocabulary task the alternatives were in Finnish. The latter task appeared to work as the constructors expected. The candidates used three strategies: if they knew the word, they went straight for the 'correct' meaning and marked it without reading the rest of the alternatives. If they were more unsure, they read through all the alternatives first and then decided which one was best. When they had no idea of the best answer, they read through the alternatives once or twice, maybe eliminated one or two of them and then just chose one using, for better or for worse, direct translation and/or what possible collocations or cognates they could think of or invent. This task, then, appears to be testing single-word knowledge; any word families or derivations that the testee uses in deciding which the best alternative is have to be supplied by the testee herself.

The test constructors were a little surprised by the apparent difficulty of some of the items in the collocations task for the Intermediate testees. Some of the items were too difficult simply because the testees did not know the usage of some/any of the alternatives, so the level of difficulty of this task needs attention. Direct translation helped the testees sometimes but sometimes also led them astray. None of the three introspectors appeared to use elimination in this task, but they did guess. The items on appropriate level made the testees search for related words and/or contexts: mother-in-law => in-laws, make one's dream come true. Here, when the items are on an appropriate difficulty level, we are testing the testees' knowledge of word usage, but we will have to pay more attention to what levels of usage knowledge we can require on these skill levels. As task types, the bi-and monolingual multiple choice items seem to work, and new types are needed only if we wish to make the whole subtest monolingual or if we want to add a test of vocabulary in context.

Pilot test data - The retrospective interviews

During the retrospective interviews after the language lab, the interviewers wanted to find out what the testees had done when performing the tasks and how they had felt about the tasks and the test as a whole. In the listening part, the testees behaved very much as expected: they listened to everything and picked up the main points or specific details as requested. Some commented that they would have felt more secure about answering some questions on bits of dialogue had they heard the texts twice. They could then have listened for gist the first time and specifically for the answer to the question the second time. As it was, they listened for overall impression and, based on that, answered the questions from memory. The questions focused on the main content of the dialogues, which made answering easier, though. Holistic understanding was indeed what the item writers had wanted to test, because the idea was to concentrate on everyday understanding. However, it can be argued that comprehension of more extended speech on more abstract topics is also an important part of language ability on these skill levels and thus deserves to be tested, at least in the Intermediate test.

None of the testees complained that the listening tasks were too easy, but those who commented on difficulty said that the speaking part was much more difficult. The facility of the listening tasks depended at least on the familiarity of the topics/situations, the

speed of the speakers and focusing on main points or specific details only. The difficulty level of the texts was approximately the same as that of the texts / contexts in the speaking part, and this level appears to be too low. Clear indications of how to make the listening part more difficult were not obtained, and further piloting is necessary.

In the speaking part, different test-taking strategies showed for example in how the testees utilized the time given for getting acquainted with each task. Some used the time for reading the tasks through carefully, some glanced the tasks through for overall impression, and some took a pause and waited for the task to begin. The biggest surprise for the test constructors was how many of the testees ignored most of the supporting strategies and hints given in the instructions, and how many of the "so clear" instructions still needed revision.

The most interesting comments in the interviews were gathered on the effect of the test mode of the speaking subtest, language laboratory. A great majority of those interviewed stated that they would rather have performed the speaking part face to face, be it then with a native or a non-native speaker. Response times were a problem, sometimes to long and sometimes too short. Uncertainty about the duration of the pauses for response also caused anxiety. Even if the instructions said there would be a 20-30 seconds pause during which to respond, the testees said it was difficult to estimate this and thus know when the pause was coming to a close; one had to hurry in order to say all one wanted, and then, after responding, one felt bad if the pause was much longer and one would have had time for a more thorough answer. On the other hand, if one answered too long, the tape cut in in the middle of a sentence and left the answer incomplete. In the simulated conversations, the missing interactivity irritated many: no matter what they said, the tape was going to continue the conversation as preprogrammed, and if they happened to produce a turn that did not fit the next turn on the tape, they felt that they had failed, and only later, if at all, thought that perhaps there was something wrong with the test.

In addition to problems with response times, many commented that having a natural intonation when talking to a machine was difficult. Also, knowing that their production was being taped all the time made the testees feel unable to correct themselves even if they noticed a flaw they had made, while when talking to a person they would automatically do that, and notice their errors more quickly, based on feedback from the conversation partner. In a natural conversation, they could also ask for a repetition, or ask clarification questions, but not here. This adds to the sense of finality. Different testees also had different abilities to imagine themselves into the situations put to them in the test. The "better actors" were able to get more intonation and naturalness to their response while those who experienced high artificiality said they just used the phrases they knew fitted in each situation, acutely feeling that all naturalness was missing. Several of the testees also said they had felt a lot of anxiety, and that they normally do better when talking face to face. A common comment was that the testees "didn't know how few words they actually knew". Two or three testees said they were totally blocked during the language lab and afterwards reported that although their language competence was not intermediate level face to face, the performance they were able to give in the language laboratory had nothing to do with what they could do in real life.

In short, the speaking part seems to be testing language knowledge more than language use and that it favours those with livelier imagination and/or those with experience on

language laboratories. All of this is more or less justified criticism against a lab-only speaking test. Based on this and other considerations, we are now developing a speaking test where one part is performed face to face. We do not want to give up the language lab totally because it has its good sides; it is economical both in the sense that many testees can be tested at once, and many different tasks can be given in a relatively short time. Some tasks, like answering short questions, reading out loud, simulated telephone conversations and monologues, suit the lab fairly well.

To sum up

The pilot testing phase proved very useful for the development and early validation of the Certificates of Foreign Language Proficiency. Robust statistical validation was practically impossible because of the small size of the samples while qualitative techniques were easily employed. The think-aloud protocols and retrospective interviews gave us important information which we could not have obtained otherwise. They allowed us to look into what our tests make the testees do when performing the test and, thus, what we are evaluating when we give grades based on test performance. This was important because we can now be a bit more certain about what it is we are testing.

Finding volunteers for the introspection part was not particularly easy, and it could be thought that those who did volunteer might not be typical representatives of the population we wished to investigate. With the kind of intensive qualitative work, volunteers were the only possibility, though. Like Alderson (1990:468) a few years ago, we contended that finding good informants is more important than finding representative informants. Volunteers for the retrospective interview were easier to find, but among them, different degrees of informativeness appeared. This appears to be a permanent characteristic of qualitative techniques and one we will have to live with.

We were very pleased with both introspection and retrospective interview as validation tools and hope that we can use them regularly when piloting new tasks. A short completing interview after the introspection, maybe after the researcher has listened to the introspection at least once, would be a good complement for this technique; it would give the researcher a chance to make sure she has interpreted some points correctly, and the testee a chance to discuss this very intensive experience. One introspecting testee is not enough, but three might be, five would be even better. Concerning our test, it would be very good to get people on different skill levels to introspect on the same test.

With the first bigger groups of testees, we hope to be able to investigate the use of more sophisticated statistical validation tools, and to compare the usefulness of qualitative and quantitative validation techniques.

REFERENCES

Alderson, Charles 1990 Testing Reading Comprehension Skills (Part Two) Getting Students to Talk About Taking a Reading Test (A Pilot Study). In Reading in a Foreign Language, 7(1), 465-503.

Bachman, Lyle F 1990 Fundamental Considerations in Language Testing. Oxford: OUP.

Davidson, Fred and David Lynch 1993 Criterion Referenced Language Test Development: A Prolegomenon. In Huhta, Sajavaara and Takala (eds) Language Testing: New Openings. Jyväskylä: Institute for Educational Research.

Kenyon, Dorry Mann and Charles W. Stansfield 1993 A Method for Improving Tasks on Performance- Based Assessments. In Huhta, Sajavaara and Takala (eds) Language Testing: New Openings. Jyväskylä: Institute for Educational Research.

Appendix 1. The Finnish 9-level Scale of Foreign Language Proficiency

9 Has a full command of the language: flawless, fluent, appropriate and well organised use of language. An exceptional level of language proficiency, which is normally attained only by well-educated language professionals.

8 Communicates effectively and appropriately even in demanding oral and written tasks and situations. Fluent and in many ways native-like. Occasional problems with subtle stylistic distinctions and idioms.

7 Communicates effectively and appropriately even in many demanding oral and written tasks and situations. Usage is quite versatile and fluent with some trace of the mother tongue. Understands with ease both general and professional/occupational language.

6 Communicates appropriately in familiar oral and written tasks and situations related to work and freetime. Language knowledge seldom hinders effective communication. Occasional inaccuracies and inadequacies which nevertheless seldom lead to misunderstandings. Mother tongue interference is in evidence but not intrusive. Rarely has problems understanding general or professional/ occupational language.

5 Communicates well in familiar oral and written tasks and situations related to work and freetime. Makes an effort to be an effective communicator. Inaccuracies cause some misunderstandings and language is not always quite fluent or appropriate. Interference from mother tongue or other languages is evident. Understands ordinary spoken and written text and there is only occasional need for repetition or consulting a dictionary.

4 Communicates fairly well on familiar tasks and situations; effective communication may sometimes be hindered by problems with language. Can handle routine writing tasks related to work and freetime. Interference from L1/other languages quite obvious. Vocabulary, grammar and fluency generally adequate, but speaking or writing may reveal specific strengths or weaknesses. A dictionary may be needed for understanding ordinary text, for instance, a newspaper article.

3 Manages to communicate in the most familiar oral and written tasks and situations but new situations cause communication problems. Understands slow and careful speech and can normally understand the gist of an easy text, for instance, a newspaper article.

2 Manages to communicate in simple and routine tasks and situations. With the help of a dictionary can understand simple written messages and without one can get the gist. Limited language proficiency causes frequent breakdowns and misunderstandings in non-routine situations.

1 Knowledge of language suffices to be able to cope with the simplest oral and written tasks and situations. Can understand the topic in newspaper articles and conversations that deal with familiar subjects. Knows some of the basic structures of the language.

This is a working version of the scale, and changes in it are possible